Feature Selection for Sentiment Classification Using Matrix Factorization

Jiguang Liang[1], Xiaofei Zhou[1], Li Guo[1], Shuo Bai[1,2]
[1]National Engineering Laboratory for Information Security Technologies
Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100190, China
[2]Shanghai Stock Exchange, Shanghai 200120, China
{liangjiguang, zhouxiaofei, guoli, baishuo}@iie.ac.cn

ABSTRACT

Feature selection is a critical task in both sentiment classification and topical text classification. However, most existing feature selection algorithms ignore a significant contextual difference between them that sentiment classification is commonly depended more on the words conveying sentiments. Based on this observation, a new feature selection method based on matrix factorization is proposed to identify the words with strong inter-sentiment distinguish-ability and intra-sentiment similarity. Furthermore, experiments show that our models require less features while still maintaining reasonable classification accuracy.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis

General Terms

Algorithms, Theory

Keywords

sentiment classification; feature selection; sentiment analysis; matrix factorization

1. INTRODUCTION

Sentiment analysis is concerned with classifying subjective text into positive or negative according to the opinions expressed in them. The dominant techniques consider sentiment classification as a binary classification problem which generally follows traditional topical text classification approaches. So there is one major difficulty: the high dimensionality of features used to capture texts. Feature selection algorithms are usually used to obtain a reduction of the original feature set by selecting most useful features for yielding better performance and less running time. However, there is a significant difference between topical and sentiment classification that the category of subjective text depends more on its component emotional words than other representative features. Nevertheless, traditional feature selection algorithms fail to take account of this point.

In this paper, from the viewpoint of the contribution of a candidate feature to distinguish sentiments, a novel feature selection method based on matrix factorization is proposed for sentiment classification. The experimental results indicate that the proposed method is effective for sentiment classification with fewer bag-of-words features.

Copyright is held by the author/owner(s).
WWW'15 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742741.

2. METHODOLOGY

One assumption that researchers often make about sentiment classification is that words that frequently appear in one category and seldom appear in the other category are more likely to have strong inter-sentiment separability [1]. To formalize this intuition, we use $D = \{d_i\}_{i=1}^m$ and $L = \{l_i\}_{i=1}^m$ to denote subjective document set and the corresponding sentiment label set. If d_i is a positive document, then $l_i = +1$; otherwise $l_i = -1$. The vocabulary index is denoted by $W = \{w_i\}_{i=1}^n$. We also consider an $m \times n$ contribution matrix R describing n words' inter-sentiment distinguish-ability on m subjective documents:

$$R_{ij} = (\mathcal{F}^{(+)}(j)/\mathcal{F}^{(-)}(j))^{l_i} \cdot \mathcal{F}^{(i)}(j)/t_i \qquad (1)$$

where $\mathcal{F}^{(i)}(j)$, $\mathcal{F}^{(+)}(j)$ and $\mathcal{F}^{(-)}(j)$ are the frequencies of w_j in d_i, positive and negative corpora. t_i is the length of d_i. Then, we can obtain a score (sentiment distinguish-ability) for each word from the perspective of the contribution to sentiment classification:

$$score(j) = AVG(R^+_{\cdot,j}) - AVG(R^-_{\cdot,j}) \qquad (2)$$

Here, $R^+_{\cdot,j}$ is the sum of R_{ij} where $l_i > 0$ and AVG is the average function. The bigger $|score(j)|$ the better inter-sentiment distinguish-ability.

However, R is a extremely sparse matrix. A low-rank matrix factorization model (MF1) is used to predict the unknown variables by minimizing

$$\begin{aligned}\min_{U,V} \mathcal{J}(R,U,V) = &\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}(R_{ij} - U_i^T V_j)^2 \\ &+ \frac{\alpha}{2}\sum_{j=1}^{n}\|V_j - \sum_{k=1}^{n} I^s_{jk} V_k\|_F^2 \\ &+ \frac{\beta}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} I^o_{jk}\|V_j + V_k\|_F^2 \\ &+ \frac{\gamma}{2}\|U\|_F^2 + \frac{\lambda}{2}\|V\|_F^2\end{aligned} \qquad (3)$$

where $\mathbf{U} \in \mathbb{R}^{l,m}$ and $\mathbf{V} \in \mathbb{R}^{l,n}$ are latent feature matrices about documents and words, $l < min(m,n)$, and $\alpha, \beta, \gamma, \lambda > 0$. I^s is similarity function and we use pointwise mutual information normalized between [0,1] to depict it. The last two regularization terms are added to avoid overfitting.

Table 1: Results in applying MF and other SVM-based methods.

Method	FeatureNum	Accuracy	Method	Assistant Information	Accuracy
IG	1800	82%	Pang & Lee, 2004	5000 subjective and 5000 objective sentences	87.15%
MI	1800	81.8%	Whitelaw, 2005	1597 appraisal groups; 48314 features	90.2%
CHI	1700	79.2%	Martineau et al.,2009	bag of words feature	88.1%
SVD	1500	87%	Maas et al., 2011	50000 additional unlabeled reviews; 5050 features	88.9%
NMF	1100	85.7%	Tu et al., 2012	part-of-speech and dependency trees	88.5%
MF1	1300	88.5%	Wang et al., 2012	NB log-count ratios; unigrams and bigrams	89.45%
MF2	1300	**89.5**%	Nguyen et al., 2013	opinion lexicons; 50000 unlabeled reviews	87.95%

The second regularization term is used to constrain similar sentiment. More specifically, two frequently co-occurring words are more likely to share similar sentiment labels. In other words, they tend to have strong intra-sentiment similarity. Then we could assume that $w_j's$ sentiment distinguish-ability should be close to the expected value of co-occurring words' distinguish-ability. However, this term is insensitive to those documents that contain words expressing both positive and negative sentiments. Hence, we propose another term to impose constraints for similar sentiments:

$$\frac{\alpha}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} I_{jk}^{s}\|V_j - V_k\|_F^2 \qquad (4)$$

The smaller I_{jk}^{s} the larger intra-sentiment similarity between w_j and w_k. This model is called MF2.

The third regularization term is to constrain antonyms. Intuitively, a pair of antonyms tend to be similar in sentiment distinguish-ability but opposite in signs (one "+" and the other "-"). We define I_{jk}^{o} as the indicator function that is equal to 1 if w_j is opposite to w_k and equal to 0 otherwise. In this paper, antonyms can be obtained by negation handling preprocess: concatenating the first word after the negation word (not, never, don't, et al.) that should not be a stop word. For example, "*not* a *good* idea" becomes to "*not_good* idea" after negation handling. Meanwhile, we can obtain a pair of antonyms "*good*" and "*not_good*".

Gradient descent algorithm is used to search the solution.

3. EXPERIMENTS

Experimental Setting: We evaluate our methods on the movie reviews dataset collected by Pang et al.[4]. We set α, β, γ and λ to 0.001, and $l = 10$. 8408 words are selected for candidate features whose document frequencies and collection frequencies are higher than 5 and 10, respectively.

Experimental Results: The best accuracy for each approach is presented in Table 1. It can be observed that our methods significantly outperform traditional feature selection methods (information gain (IG), Chi-square statistics (CHI) and mutual information (MI)). Besides, our methods with 1300 features are better than or comparable to previous works using much more unlabeled data, features and priori information which are often expensive to obtain. Whitelaw et al.[7] got the best accuracy 90.2%. However, this method is very complicated using 1597 appraisal groups and 48314 features. A detail analysis about the effects of feature number (FN) to accuracy is shown in Fig 1 from which we can find that our methods could produce effective and stable results (>88.6%) when FN >1000.

Case Study: Besides, our models' top scoring features are clearly more sentimental than baselines. Consider the example in Table 2. Our models could place much greater weight on words that convey sentiments than objective words.

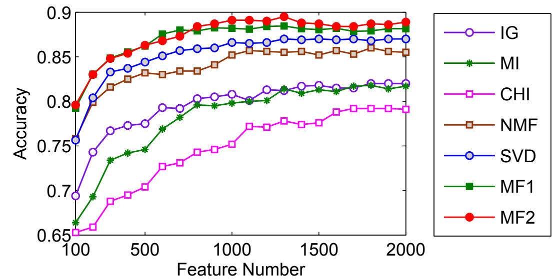

Figure 1: Effects of feature number to accuracy.

Table 2: Top-5 features for negative corpus.

IG	NMF	SVD	MF1	MF2
film	seagal	seagal	mulan	**bad**
his	brenner	brenner	seagal	**worst**
it's	general's	**bad**	lebowski	jawbreaker
movie	wayans	movie	**bad**	**stupid**
life	**bad**	general's	**worst**	**boring**

4. CONCLUSIONS

In this paper, we introduce a matrix factorization framework for sentiment feature selection. Experimental results show that our models outperform most published results on Movie dataset.

Acknowledgments

This work was supported by Strategic Priority Research Program of Chinese Academy of Sciences (XDA06030600) and National Nature Science Foundation of China (No.61202226).

5. REFERENCES

[1] Liang J.G., Zhou X.F., Hu Y., et al. CONR: A Novel Method for Sentiment Word Identification. CIKM, 2014.

[2] Maas A L, Daly R E, Pham P T, et al. Learning word vectors for sentiment analysis. ACL, 142-150, 2011.

[3] Martineau J, Finin T. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. ICWSM, 2009.

[4] Nguyen D Q, Nguyen D Q, Pham S B. A two-stage classifier for sentiment analysis. IJCNLP, 2013.

[5] Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. ACL, 2004.

[6] Tu Z., He Y., et al. Identifying high-impact sub-struc -tures for convolution kernels in document-level senti -ment classification. ACL, 338-343, 2012.

[7] Wang S., Manning C.D. Baselines and bigrams: simple, good sentiment and topic classification. ACL, 2012.

[8] Whitelaw C, Garg N, Argamon S. Using appraisal groups for sentiment analysis. CIKM, 625-631, 2005.